An Introduction to Sequential Pattern Mining

Posted on 2017-03-08 by Philippe Fournier-Viger

In this blog post, I will give an introduction to **sequential pattern mining,** an important data mining task with a wide range of applications from text analysis to market basket analysis. This blog post is aimed to be a short introductino. If you want to read a more detailed introduction to **sequential pattern mining**, you can read a survey paper that I recently wrote on this topic.

What is sequential pattern mining?

Data mining consists of extracting information from data stored in databases to understand the data and/or take decisions. Some of the most fundamental data mining tasks are clustering, classification, outlier analysis, and pattern mining. **Pattern mining** consists of discovering interesting, useful, and unexpected patterns in databases Various types of patterns can be discovered in databases such as frequent itemsets, associations, subgraphs, sequential rules, and periodic patterns.

The task of **sequential pattern mining** is a data mining task specialized for analyzing **sequential data,** to discover **sequential patterns**. More precisely, it consists of discovering interesting subsequences in **a set of sequences**, where the interestingness of a subsequence can be measured in terms of various criteria such as its occurrence frequency, length, and profit. Sequential pattern mining has numerous real-life applications due to the fact that data is naturally encoded as **sequences of symbols** in many fields such as bioinformatics, e-learning, market basket analysis, texts, and webpage click-stream analysis.

I will now explain the task of **sequential pattern mining** with an example. Consider the following **sequence database**, representing the purchases made by customers in a retail store.

SID	Sequence
1	$\langle\{a,b\},\{c\},\{f,g\},\{g\},\{e\}\rangle$
2	$\langle\{a,d\},\{c\},\{b\},\{a,b,e,f\}\rangle$
3	$\langle\{a\},\{b\},\{f,g\},\{e\}$
4	$\langle\{b\},\{f,g\}\rangle$

This database contains four sequences. Each **sequence** represents the items purchased by a customer at different times. A sequence is an ordered list of itemsets (sets of items bought together). For example, in this database, the first sequence (SID 1) indicates that a customer bought some items *a* and *b* together, then purchased an item *c*, then purchased items *f* and *g* together, then purchased an item *g*, and then finally purchased an item *e.*

Traditionally, sequential pattern mining is being used to find subsequences that appear often in a sequence database, i.e. that are common to several sequences. Those subsequences are called the **frequent sequential patterns**. For example, in the context of our example, sequential pattern mining can be used to find the sequences of items frequently bought by customers. This can be useful to understand the behavior of customers to take marketing decisions.

To do **sequential pattern mining**, a user must provide a sequence database and specify a parameter called the **minimum support threshold**. This parameter indicates a minimum number of sequences in which a pattern must appear to be considered frequent, and be shown to the user. For example, if a user sets the minimum support threshold to 2 sequences, the task of **sequential pattern mining** consists of finding all subsequences appearing in at least 2 sequences of the input database. In the example database, 29 subsequences met this requirement. These sequential patterns are shown in the table below, where the number of sequences containing each pattern (called the *support*) is indicated in the right column of the table.

Pattern	Sup.
$\langle\{a\}\rangle$	3
$\langle\{a\},\{g\}\rangle$	2
$\langle\{a\},\{g\},\{e\}\rangle$	2
$\langle\{a\},\{f\}\rangle$	3
$\langle\{a\},\{f\},\{e\}\rangle$	2
$\langle\{a\},\{c\}\rangle$	2
$\langle\{a\},\{c\},\{f\}\rangle$	2
$\langle\{a\},\{c\},\{e\}\rangle$	2
$\langle\{a\},\{b\}\rangle$	2
$\langle\{a\},\{b\},\{f\}\rangle$	2
$\langle\{a\},\{b\},\{e\}\rangle$	2
$\langle\{a\},\{e\}\rangle$	3
$\langle\{a,b\}\rangle$	2
$\langle\{b\}\rangle$	4
$\langle\{b\},\{g\}\rangle$	3
$\langle\{b\},\{g\},\{e\}\rangle$	2
$\langle\{b\},\{f\}\rangle$	4
$\langle\{b\},\{f,g\}\rangle$	3
$\langle\{b\},\{f\},\{e\}\rangle$	2
$\langle\{b\},\{e\}\rangle$	3
$\langle\{c\}\rangle$	2
$\langle\{c\},\{f\}\rangle$	2
$\langle\{c\},\{e\}\rangle$	2
$\langle\{e\}\rangle$	3
$\langle\{f\}\rangle$	4
$\langle\{f,g\}\rangle$	3
$\langle\{f\},\{e\}\rangle$	2
$\langle\{g\}\rangle$	3
$\langle\{g\},\{e\}\rangle$	2

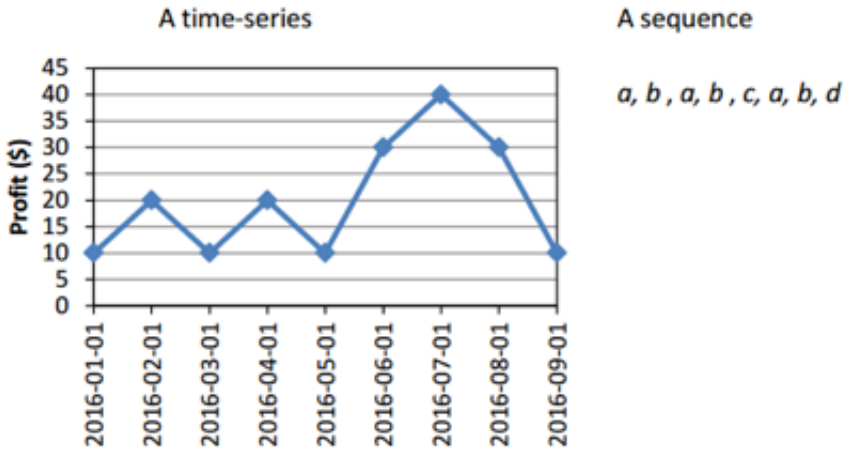
For example, the patterns <{a}> and <{a}, {g}> are frequent and have a support of 3 and 2 sequences, respectively. In other words, these patterns appears in 3 and 2 sequences of the input database, respectively. The pattern <{a}> appears in the sequences 1, 2 and 3, while the pattern <{a}, {g}> appears in sequences 1

and 3. These patterns are interesting as they represent some behavior common to several customers. Of course, this is a toy example. Sequential pattern mining can actually be applied on database containing hundreds of thousands of sequences.

Another example of application of sequential pattern mining is text analysis. In this context, a set of sentences from a text can be viewed as sequence database, and the goal of sequential pattern mining is then to find subsequences of words frequently used in the text. If such sequences are contiguous, they are called "ngrams" in this context. If you want to know more about this application, you can read this blog post, where sequential patterns are discovered in a Sherlock Holmes novel.

Can sequential pattern mining be applied to time series?

Besides sequences, **sequential pattern mining** can also be applied to **time series** (e.g. stock data), when discretization is performed as a pre-processing step. For example, the figure below shows a **time series** (an ordered list of numbers) on the left. On the right, a **sequence** (a sequence of symbols) is shown representing the same data, after applying a transformation. Various transformations can be done to transform a time series to a sequence such as the popular SAX transformation. After performing the transformation, any sequential pattern mining algorithm can be applied.



A time-series (left) and a sequence (right)

Where can I get Sequential pattern mining implementations?

To try sequential pattern mining with your datasets, you may try the open-source **SPMF data mining software**, which provides implementations of numerous **sequential pattern mining algorithms:** http://www.philippe-fournier-viger.com/spmf/

It provides implementations of several algorithms for sequential pattern mining, as well as several variations of the problem such as discovering **maximal sequential patterns**, **closed sequential patterns** and sequential rules. Sequential rules are especially useful for the purpose of performing predictions, as they also include the concept of confidence.

What are the current best algorithms for sequential pattern mining?

There exists several sequential pattern mining algorithms. Some of the classic algorithms for this problem are **PrefixSpan, Spade, SPAM,** and **GSP**. However, in the recent decade, several novel and more efficient algorithms have been proposed such as **CM-SPADE** and **CM-SPAM** (2014), **FCloSM** and **FGenSM** (2017), to name a few. Besides, numerous algorithms have been proposed for extensions of the problem of sequential pattern mining such as finding the sequential patterns that generate the most profit (high utility sequential pattern mining).

Conclusion

In this blog post, I have given a brief overview of **sequential pattern mining**, a very useful set of techniques for analyzing sequential data. If you want to know more about this topic, you may read the following recent survey paper that I wrote, which gives an easy-to-read overview of this topic, including the algorithms forf sequential pattern mining, extensions, research challenges and opportunities.

Fournier-Viger, P., Lin, J. C.-W., Kiran, R. U., Koh, Y. S., Thomas, R. (2017). **A Survey**

of Sequential Pattern Mining. Data Science and Pattern Recognition, vol. 1(1), pp. 54-77.

—

Philippe Fournier-Viger is a professor of Computer Science and also the founder of the open-source data mining software SPMF, offering more than 120 data mining algorithms.